Manifold Learning for ToF-based Human Body Tracking and Activity Recognition

Loren Arthur Schwarz
schwarz@cs.tum.edu
Diana Mateus
mateus@cs.tum.edu
Victor Castañeda
castaned@cs.tum.edu
Nassir Navab
navab@cs.tum.edu

Chair for Computer Aided Medical Procedures & Augmented Reality
Technische Universität München
Garching bei München, Germany

Abstract

In this paper, we propose a method for simultaneous human full-body pose tracking and activity recognition from time-of-flight (ToF) camera images. Simple and sparse depth cues are used together with a prior motion model that constrains the tracking problem. Our model consists of low-dimensional manifolds of feasible poses for multiple activities. A particle filter allows us to efficiently evaluate various pose hypotheses over different activities and to select one that is most consistent with the observed depth image cues. We relate poses in the manifold embeddings to full-body poses and to observable depth cues using non-linear regression mappings. Our method is able to robustly detect changes of activity and adapt accordingly. We evaluate our method on a dataset containing 10 activities for 10 persons and show that we can track full-body pose and classify performed activities with a high precision which is discussed in the paper.

1 Introduction

Recent technological advances have lead to the development of cameras that measure depth by means of the time-of-flight (ToF) principle [15]. ToF cameras allow capturing an entire scene instantaneously, and thus provide depth images in real-time. Despite the relatively low resolution, this type of data offers a clear advantage over conventional cameras for some applications. Human-machine interaction is an example where real-time is required and where depth information is a valuable cue. In this paper, we investigate the use of depth cues from ToF cameras for interactions governed by human motion. Specifically, we propose a method both to *recognize* the current activity and to *track* the full-body pose of a person observed by a ToF camera. Simultaneous recognition and tracking allows for rich interactions, since the exact pose and variations, e.g. the speed of execution, can be taken into account.

Over the past years, action recognition and pose estimation from monocular videos [10, 17, 23] have received a lot of attention. These are difficult problems since the 2D projection of human movements is, by the nature of the imaging process, prone to ambiguities. The problem can be simplified if depth cues are provided [26]. The use of 3D data has been

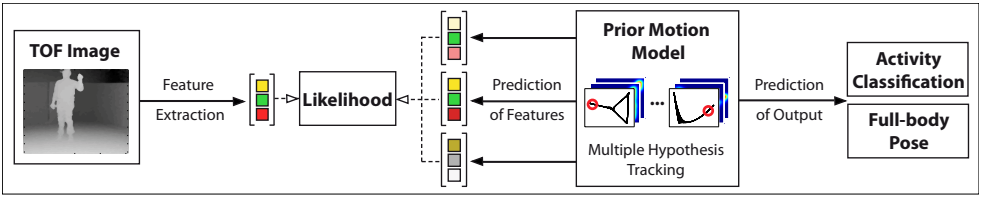

Figure 1: Overview of the proposed full-body tracking and activity recognition method. The learned prior motion model is based on low-dimensional, activity-specific manifolds of feasible poses. A particle filter is used to sample the reduced pose space. The highest-scoring hypothesis is used for activity classification and full-body pose estimation.

studied in [25] for action recognition and in [19] for pose estimation using a multi-view camera system and 3D laser scanners, respectively. These types of observations are difficult to obtain in real-time and require complex hardware. ToF cameras avoid such problems and permit building systems that are suitable for 3D interactions [2, 21]. However, ToF data is sparse and suffers from noise and motion blur, and gives rise to other challenges in action recognition and especially full-body tracking.

We propose an efficient method for activity recognition and body tracking based on simple depth features, a strong prior motion model and a sampling-based inference approach (Figure 1). Our motion model removes the need for fitting a skeleton using computationally expensive optimisation techniques, as in [2], and allows us to rely on easily obtainable depth features for tracking. To prevent being dependent on error prone body part detection in ToF data, we design a global feature representation describing the human body shape at each instant. As opposed to feature descriptors commonly used for 3D shape recognition [7, 12], our feature representation varies continuously with the performed motion.

The central component of our method is the prior motion model that is based on a set of low-dimensional *manifold embeddings* for each activity of interest. We apply a manifold learning technique [3] to generate the embeddings from full-body pose training data. Each of the embeddings acts as a low-dimensional parametrisation of feasible body poses [4] that we use to constrain the tracking problem. Instead of exhaustively searching the high-dimensional full-body pose space, we deploy a particle filter in the low-dimensional manifold embedding space. This way, our method is able to track multiple pose hypotheses for different activities and to select one that is most consistent with the depth cue observations. In order to link the manifold embeddings to full-body poses (joint angles) and observations (depth features), we learn *predictive mappings* by means of non-linear regression. Our approach combines the distinctiveness of multiple local, activity-specific motion models into a global model capable of recognising and tracking multiple activities from simple observations. The results provided in the evaluation section demonstrate accurate activity recognition and full-body tracking for 10 persons and 10 different activities.

1.1 Related Work

Recently, several approaches for human action (or gesture) recognition using ToF images have been proposed. For instance, Penne *et al*. use a ToF camera for hand gesture recognition with the aim of manipulating a medical 3D dataset [16, 21]. The method is based on a classifier trained to distinguish the surface appearance of a set of pre-defined hand poses.

Holte *et al*. describe a technique for recognizing upper-body gestures, such as raising one or both arms, from ToF camera images [7]. Jensen *et al*. propose an approach for gait tracking using whole-body ToF images [11]. The method is based on localising leg joint positions in range images of a person walking on a treadmill. Zhu *et al*. present a technique for upper-body tracking by fitting a body model to the ToF data, after identification of anatomical landmarks [27]. A more general, whole-body tracking approach is described by Plagemann *et al*. [6, 18]. Their method detects interest points in each ToF image and associates them with hands, head and feet of a body model. As opposed to separately targeting recognition [7, 16, 21] and tracking [6, 11, 27], we perform the two tasks simultaneously. Moreover, the aforementioned approaches for pose tracking are highly dependent on detecting body parts in the ToF data and are therefore susceptible to noise and self-occlusions. Another strategy for feature extraction is to use general 3D shape descriptors, such as shape contexts or spin images [12]. This type of features have been used for gesture recognition from ToF data [7]. However, these descriptors are mainly suitable for shape classification and do not adequately represent subtle pose changes, as required for full-body tracking. We design a simple but general feature descriptor that, first, does not rely on body-part detection and, second, varies smoothly with the movements of a person. In fact, it is our prior motion model that enables full-body tracking from such simple depth cues.

Prior models have been used for constraining the tracking problem from observations such as silhouettes in monocular videos [10, 23] or wearable sensor data [20]. The common idea is to avoid searching the high-dimensional full-body pose space and to use a learned parametrisation of feasible human poses instead. Several authors have proposed tracking methods based on the Gaussian Process Latent Variable Model (GPLVM) [22, 23], where a low-dimensional latent space of poses for a given activity is learned from training data. Other authors use manifold learning techniques, such as Isomap [10] or Laplacian Eigenmaps [17], for obtaining low-dimensional pose priors. We also choose a manifold learning method over GPLVMs since the latter are computationally significantly more expensive. In addition, manifold embeddings have the favourable property of preserving the local spatial relationships of the high-dimensional input data.

2 Method

We propose a method for full-body pose tracking and activity recognition from simple depth cues. The problem is constrained by means of a prior motion model learned during a training phase from full-body pose data (Figure 1). Our model is based on activity-specific manifold embeddings that can be seen as an independent, low-dimensional parametrisation of feasible poses for each activity of interest.

Formally, let $\mathbf{y} \in \mathbb{R}^{d_y}$ denote a full-body pose, consisting of the joint angles of our skeleton model and let $\mathbf{s} \in \mathbb{R}^{d_s}$ be a feature vector representing a ToF depth image (section 2.1). We are given a training dataset of labelled full-body poses and ToF feature vectors $\{\mathbf{Y}^\alpha, \mathbf{S}^\alpha\}$, $\alpha \in \{1,\ldots,M\}$, for M activities of interest. Each activity α contains N_α training poses, i.e. $\mathbf{Y}^\alpha = [\mathbf{y}_1^\alpha,\ldots,\mathbf{y}_{N_\alpha}^\alpha]$ and $\mathbf{S}^\alpha = [\mathbf{s}_1^\alpha,\ldots,\mathbf{s}_{N_\alpha}^\alpha]$. During the training phase, our objective is to learn a prior motion model based on manifold embeddings $\mathbf{X}^\alpha = [\mathbf{x}_1^\alpha,\ldots,\mathbf{x}_{N_\alpha}^\alpha]$ for each activity (section 2.2). The link from positions in embedding space to full-body poses and feature vectors is created by means of two predictive mappings learned from training data using non-linear regression. We also introduce prior knowledge for keeping pose predictions close to the poses in the training dataset and for modelling the activity switching process.

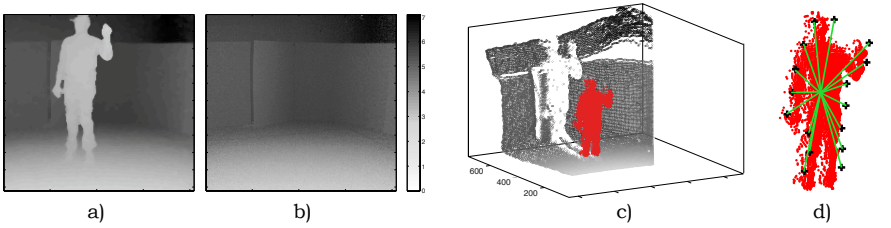

Figure 2: ToF feature extraction process. a) Filtered depth image, b) background image, c) 3D segmented foreground, d) feature descriptor based on 3D vectors (lines) from the centroid of the segmented person to extremal points on the surface boundary (crosses).

In the testing phase, we wish to recognize the performed activity $\hat{\alpha}_t$ and to predict the full-body pose $\hat{\mathbf{y}}_t$ at every time step t, given only observed feature vectors $\mathbf{s}_t$. We model the state of our dynamic system as a pair $(\hat{\alpha}_t, \hat{\mathbf{x}}_t)$ of an activity index and a position in the corresponding manifold embedding. For state inference, we employ a particle filter that efficiently samples the embedding space and tracks multiple pose hypotheses (section 2.3). The features, the prior motion model and the inference approach are described in the following.

2.1 Feature Extraction from ToF Images

Given a ToF depth map $\mathbf{I}_t$ at time t, we obtain a feature vector $\mathbf{s}_t \in \mathbb{R}^{d_s}$ in three steps: (1) preprocessing of the image to remove noise and to convert the depth map to 3D information, (2) segmentation of the person from the scene background and (3) extraction of simple features related to the boundaries of the observed 3D point cloud. Initially, we apply a median filter to the depth map to remove noise. We then transform the filtered depth map to a 3D point cloud using the intrinsic parameters of the ToF camera [6]. This representation of the scene, denoted by $\tilde{\mathbf{I}}_t$, is invariant to scaling caused e.g. by the person moving towards the camera. In order to segment a person in front of the ToF camera, we perform static background subtraction [14]. This process is illustrated in Figure 2. We discard all 3D points in $\tilde{\mathbf{I}}_t$ if their z coordinate is close to that of the corresponding points in the background model $\tilde{\mathbf{I}}_{\text{bg}}$. The centroid of the remaining 3D points is computed and the 3D space occupied by the points is divided into b cells, followed by determining the bounding box of each cell. We then extract the vectors $\mathbf{v}_k$, $k \in \{1,\ldots,d\}$, connecting the centroid to the bounding box corners that are furthest apart and closest to the camera. These corner points are equivalent to a sparse silhouette representation but in 3D. The feature descriptor is then given by $\mathbf{s}_t = [\mathbf{v}_1 \ldots \mathbf{v}_b]^\top$. For noise reduction, we apply a moving average that includes the previous feature vector $\mathbf{s}_{t-1}$.

2.2 Low-dimensional Prior Motion Model

The prior motion model consists of the following components that are learned for each activity: (1) a low-dimensional manifold embedding of feasible poses, (2) predictive mappings from embedding space to full-body space and to feature space, (3) a pose likelihood prior and (4) an activity switching prior. These components are now described in more detail.

Manifold Embeddings. For each activity $\alpha \in \{1,\ldots,M\}$, we learn a *manifold embedding* $\mathbf{X}^\alpha = [\mathbf{x}_1^\alpha,\ldots,\mathbf{x}_{N_\alpha}^\alpha]$ from the full-body pose training data $\mathbf{Y}^\alpha$. Each embedding point $\mathbf{x}_i^\alpha \in \mathbb{R}^{d_x}$ corresponds to a full-body pose $\mathbf{y}_i^\alpha \in \mathbb{R}^{d_y}$ and $d_x \ll d_y$. The full-body pose rep-

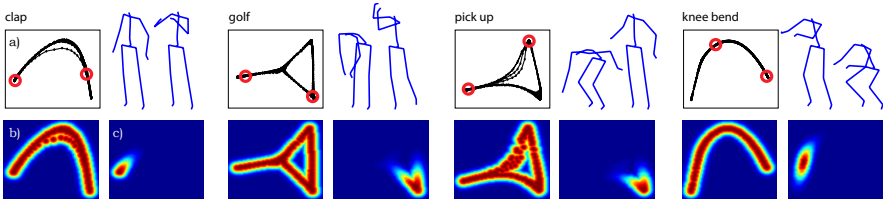

Figure 3: Learned motion models for 4 activities. a) Manifold embeddings obtained from full-body pose training data. Each point on the manifolds (left) corresponds to a full-body pose (right). Static pose priors (b) and activity switching priors (c) in embedding space.

resentation is based on our skeleton model with $d_y = 35$ degrees of freedom. We generate the manifold embeddings using Laplacian Eigenmaps [3]. Our experiments have shown that 2D embeddings are already discriminative enough to represent different poses of one activity. Increasing the dimensionality thus mainly affects computational efficiency. Note that the system state is not required to coincide exactly with the known embedding points $\mathbf{x}_i^\alpha$, allowing us to track poses that differ from the training data.

Predictive Mappings. In order to predict full-body poses $\hat{\mathbf{y}}$ and depth feature vectors $\hat{\mathbf{s}}$ from given manifold embedding positions $\mathbf{x}$, we define the *predictive mappings* $f_{xy}^\alpha : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}$ and $f_{xs}^\alpha : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_s}$ for each activity α. Following the approach in [13, 17], we use non-linear kernel regression to model these mappings as

$$\hat{\mathbf{y}} = f_{xy}^\alpha(\mathbf{x}) = \sum_{i=1}^{N_\alpha} \frac{k(\mathbf{x}, \mathbf{x}_i^\alpha)}{\sum_{j=1}^{N_\alpha} k(\mathbf{x}, \mathbf{x}_j^\alpha)} \mathbf{y}_i^\alpha \qquad \text{and} \qquad \hat{\mathbf{s}} = f_{xs}^\alpha(\mathbf{x}) = \sum_{i=1}^{N_\alpha} \frac{k(\mathbf{x}, \mathbf{x}_i^\alpha)}{\sum_{j=1}^{N_\alpha} k(\mathbf{x}, \mathbf{x}_j^\alpha)} \mathbf{s}_i^\alpha, \qquad (1)$$

where $k(\cdot,\cdot)$ is a Gaussian kernel function with a width that we determine from the standard deviation of the embedding points $\mathbf{x}_i^\alpha$. Intuitively, the mappings compute a weighted average of the full-body poses $\mathbf{y}_i^\alpha$ and feature vectors $\mathbf{s}_i^\alpha$ in the training data, with weights proportional to the similarity of the embedding location $\mathbf{x}$ to the embedding points $\mathbf{x}_i^\alpha$.

Pose Likelihood Priors. The *pose likelihood prior* $p_{\text{pose}}(\alpha, \mathbf{x})$ gives the probability that an embedding space position $\mathbf{x}$ represents a valid human pose of activity α. Intuitively, we want this probability to be high for locations close to the embedding points $\mathbf{x}_i^\alpha$. Figure 3 gives examples of pose likelihood priors. We define the prior as a *distance transform* in manifold embedding space, with a distance measure given by $k(\cdot,\cdot)$: $p_{\text{pose}}(\alpha, \mathbf{x}) = \max_{i \in \{1,\ldots,N_\alpha\}} k(\mathbf{x}, \mathbf{x}_i^\alpha)$. This expression is essentially equivalent to a kernel density estimate (KDE), with the summation in the KDE being replaced by a maximum. This ensures that the prior probability does not increase if a pose is repeated more than others (e.g. due to very slow movements).

Activity Switching Priors. The *activity switching prior* $p_{\text{switch}}(\alpha, \mathbf{x})$ describes how likely an activity switch is at a location $\mathbf{x}$ on the embedding of activity α. To ensure generality, we allow activity switching from any pose with constant minimum probability p_k. However, we let the probability of switching increase for poses that typically occur between subsequent activities. In our experiments, the upright standing pose was used as an intermediate pose. We model the switching prior with a normal distribution $p_{\text{switch}}(\alpha, \mathbf{x}) = \mathcal{N}(f_{xy}^\alpha(\mathbf{x}); \mathbf{y}_0, \Sigma_y^\alpha) + p_k$, where $\mathbf{y}_0$ represents the intermediate pose in full-body space, $f_{xy}^\alpha(\mathbf{x})$ is a predicted pose and Σ_y^α is the diagonal covariance matrix of the training data $\mathbf{Y}^\alpha$.

2.3 Pose Tracking and Activity Recognition

We infer the state $(\hat{\alpha}_t, \hat{\mathbf{x}}_t)$ of our dynamic system at each time step by means of a particle filter [8, 9] that allows estimating the posterior density of states given the observed features. The algorithm updates the particles iteratively following a *dynamics model* and an *observation model*. We initialize n particles $\mathbf{p}_0^i = (\alpha_0^i, \mathbf{x}_0^i)$, $i \in \{1, \ldots, n\}$, with locations across all manifold embeddings. Initially, the particle locations are randomly chosen according to the probabilities given by the pose likelihood priors $p_{\text{pose}}(\alpha, \mathbf{x})$. The weights w_0^i associated with the particles are initialized uniformly. Then, for each iteration t, we perform the following steps: (1) Given the particles and weights $\{\mathbf{p}_{t-1}^i, w_{t-1}^i\}$ from time $t-1$, select particles with a probability proportional to their weight, to obtain the new set $\{\bar{\mathbf{p}}_{t-1}^i, \bar{w}_{t-1}^i\}$. (2) Compute a prediction $\mathbf{p}_t^i = (\mathbf{x}_t^i, \alpha_t^i)$ for each particle by sampling from the dynamics model $p(\mathbf{x}_t, \alpha_t | \bar{\mathbf{x}}_{t-1}^i, \bar{\alpha}_{t-1}^i)$. (3) Evaluate each particle $\mathbf{p}_t^i$ against the current observation $\mathbf{s}_t$ by computing its weight $w_t^i = p(\mathbf{s}_t | \mathbf{x}_t^i, \alpha_t^i)$ using the observation model.

Dynamics Model. The dynamics model governs the evolution of particles through state space. It represents the probability distribution for new pose hypotheses, based on a previous particle state, $\bar{\mathbf{p}}_{t-1}^i = (\bar{\alpha}_{t-1}^i, \bar{\mathbf{x}}_{t-1}^i)$. Following [9], we define it to be a product of two terms:

$$p(\mathbf{x}_t, \alpha_t | \bar{\mathbf{x}}_{t-1}^i, \bar{\alpha}_{t-1}^i) = p(\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}^i, \alpha_t, \bar{\alpha}_{t-1}^i)\; p(\alpha_t | \bar{\mathbf{x}}_{t-1}^i, \bar{\alpha}_{t-1}^i), \tag{2}$$

$$p(\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}^i, \alpha_t, \bar{\alpha}_{t-1}^i) = \begin{cases} p(\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}^i) & \text{if } \alpha_t = \bar{\alpha}_{t-1}^i, \\ p_{\text{pose}}(\alpha_t, \mathbf{x}_t) & \text{else.} \end{cases} \tag{3}$$

$$p(\alpha_t | \bar{\mathbf{x}}_{t-1}^i, \bar{\alpha}_{t-1}^i) = \begin{cases} 1 & \text{if } \alpha_t = \bar{\alpha}_{t-1}^i, \\ p_{\text{switch}}(\bar{\alpha}_{t-1}^i, \bar{\mathbf{x}}_{t-1}^i) & \text{else.} \end{cases} \tag{4}$$

The pose dynamics model (Equation 3) describes particle movement through the embedding space. When no activity switch occurs, we define it to be a random walk $p(\mathbf{x}_t | \bar{\mathbf{x}}_{t-1}^i) = \mathcal{N}(\mathbf{x}_t; \bar{\mathbf{x}}_{t-1}^i, \Sigma_x^{\alpha_t})$, otherwise, p_{pose} is used. Here, $\Sigma_x^{\alpha_t}$ is the diagonal covariance matrix of the manifold embedding points $\mathbf{X}^{\alpha_t}$. The activity transition model (Equation 4) describes the activity switching process. It is based on the assumption that switching to all activities is equally likely and that the probability of an activity switch only depends on the previous particle location $\bar{\mathbf{x}}_{t-1}^i$ in the embedding space of activity $\bar{\alpha}_{t-1}^i$.

The prediction step for a particle from $\bar{\mathbf{p}}_{t-1}^i$ to $\mathbf{p}_t^i$ consists of the following steps. We sample the activity transition model (Equation 4) by selecting a new activity index α_t^i. With a probability proportional to $p_{\text{switch}}(\bar{\alpha}_{t-1}^i, \bar{\mathbf{x}}_{t-1}^i)$, this activity is randomly chosen such that $\alpha_t^i \neq \bar{\alpha}_{t-1}^i$. We determine the new particle location $\mathbf{x}_t^i$ by sampling the pose dynamics model (Equation 3). The new location is a random walk from the previous position, $\bar{\mathbf{x}}_{t-1}^i$, if $\alpha_t^i = \bar{\alpha}_{t-1}^i$. Otherwise, the particle is placed on the manifold embedding of the new activity, following the pose likelihood prior that favours locations close to the training poses.

Observation Model. The observation model allows evaluating how consistent the pose hypothesis represented by a particle is with the observed depth features. We define the observation model as a product of three terms – a prediction term, a pose smoothness term and the pose likelihood prior:

$$p(\mathbf{s}_t | \mathbf{x}_t^i, \alpha_t^i) = \mathcal{N}(\mathbf{s}_t; f_{xs}^{\alpha_t^i}(\mathbf{x}_t^i), \Sigma_s^{\alpha_t^i})\;\; \mathcal{N}(\mathbf{y_{t-1}}; f_{xy}^{\alpha_t^i}(\mathbf{x}_t^i), \Sigma_y^{\alpha_t^i})\;\; p_{\text{pose}}(\alpha_t^i, \mathbf{x}_t^i). \tag{5}$$

The prediction term uses the learned mapping $f_{xs}^{\alpha}(\mathbf{x})$ to predict depth feature vectors from an embedding position $\mathbf{x}_t^i$. This term is maximal if the prediction perfectly matches the

true observation $\mathbf{s}_t$. To reduce the influence of outlier observations, the smoothness term penalizes pose hypotheses if their predicted full-body pose differs strongly form the previous pose $\mathbf{y}_{t-1}$. Σ_s^α and Σ_y^α are the diagonal covariance matrices of the training observations and full-body poses for the respective activity.

State Estimation. We determine the system state at every time instant t as follows: The estimated activity $\hat{\alpha}_t$ is selected as the most frequent activity among the k particles with the highest weights. The pose estimate $\hat{\mathbf{x}}_t$ in manifold embedding space is computed as a convex combination of the positions of the highest-weight particles with activity $\hat{\alpha}_t$. Finally, we predict the full-body pose as $\hat{\mathbf{y}}_t = f_{xy}^{\hat{\alpha}_t}(\hat{\mathbf{x}}_t)$.

3 Experiments and Results

The following sections present the results of our evaluation. Since a synchronized dataset of ToF images and motion capture data is not available online, we recorded a database using a PMDVision CamCube ToF camera (204×204 pixels resolution) and an ART Dtrack2 tracking system. Depth features were extracted from the ToF images by subdividing the segmented 3D point cloud into 16 cells (8 vertical, 2 horizontal). The feature descriptor thus has $d_s = 16 \times 3 = 48$ dimensions. We generated manifold embeddings of $d_x = 2$ dimensions.

We considered $M = 10$ activities: clapping, golfing, hurrah (arms up), jumping jack, knee bends, picking something up, punching, scratching head, playing the violin and waving. Each of the movements was recorded 6 times with 10 actors. The testing data consists of 6 sequences per actor containing all activities in a row ($\sim$ 1500 frames per sequence). Only the depth features were used for testing, the motion capture data served as ground truth. Our experiments on classification (section 3.2) and pose estimation (section 3.3) were performed in a cross-validation scheme, i.e. each testing sequence was generated from one of the recordings per activity and actor, using the remaining five for training.

3.1 Pose Tracking

All presented experiments have been performed with $n = 300$ particles. The appropriate number of particles grows linearly with the number of considered activities. Using 300 particles, sufficiently many particles can sample the 10 embeddings, while keeping the computational complexity reasonable (our MATLAB code runs at 1 frame per second).

Figure 4 illustrates the behaviour of the particle filter on a sample sequence of two activities: *pickup* and *golfing*. The figure shows two out of the 10 used manifold embeddings. When the person starts leaning forward for picking something up (frame 33), the number of particles sampling the pickup manifold quickly increases and remains at a high level ($\sim$ 200 particles) until the person returns to the standing pose (frame 135). The number of particles sampling a particular manifold embedding is a measure of the algorithm's certainty about an activity classification. A certain number of particles constantly samples all other manifold embeddings. With the onset of the *golfing* move (frame 174), the particles on the golfing manifold are attributed higher weights and take overhand. In practice, the particle filter allows to robustly detect activity switches and thus to select the most suitable prior model for pose estimation.

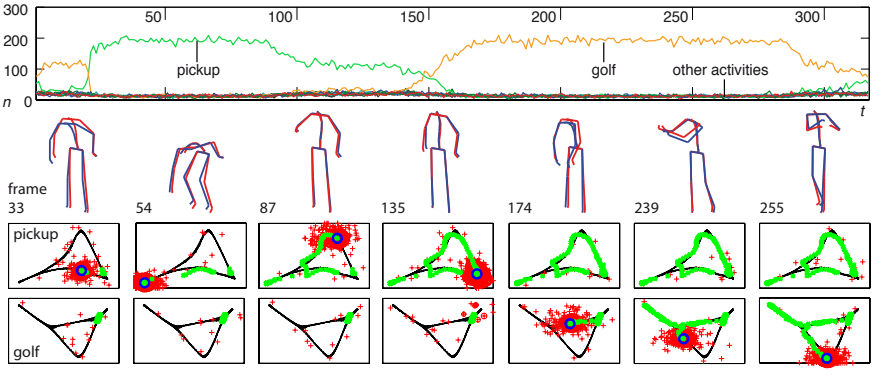

Figure 4: Particle behaviour over a sequence of 320 frames changing from *pickup* to *golf*. *Top:* Number of particles (n) on each of the 10 activity-specific manifold embeddings. Most particles sample two of the embeddings. *Middle:* Predicted (red) and ground truth (blue) full-body poses for selected frames of the sequence. *Bottom:* Manifold embeddings of pickup and golf. Particles (red), current state (dark blue) and state trace (green) are displayed.

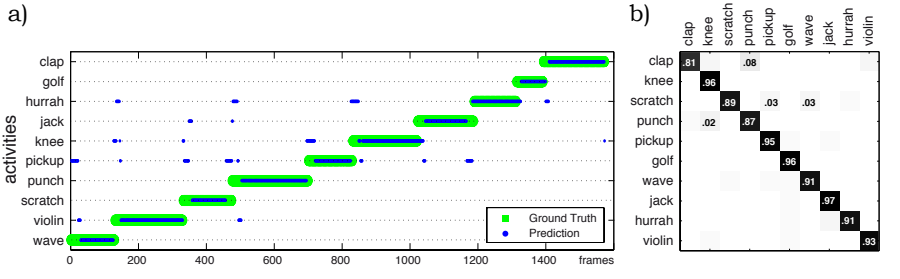

Figure 5: (a) Activity classification results for one of the testing sequences. Ground truth classification (green) and predicted activities (blue) are shown for each frame of the sequence. (b) Confusion matrix of the classification results for all testing sequences.

3.2 Activity Classification

Figure 5.(a) shows activity classification results for *one* of the testing sequences. Misclassifications mainly occur at the beginning and end of activities which correspond to the idle standing pose common to all activities. The confusion matrix in Figure 5.(b) gives the classification rates for all activities over *all* testing sequences. On average, we achieved a correct classification rate of 92% for all non-idle frames. The matrix is mostly diagonal, minor confusion only occurs between activities that consist of similar full-body poses, such as *waving* and *scratching head*. Misclassification in these cases therefore does not necessarily affect the precision of full-body pose estimation.

3.3 Pose Estimation Accuracy

We measured how precisely the poses estimated by our method match the ground truth using two metrics. The angular error e_{ang} gives the deviation from the ground truth in terms of joint angles. The distance error e_{dist} is the difference in 3D space between predicted joint locations

	clap	golf	hurrah	jack	knee	pickup	punch	scratch	violin	wave
e_{ang}	3.00	4.02	5.47	8.78	4.64	3.51	3.67	2.56	3.64	2.83
σ_{ang}	1.90	1.95	3.62	4.26	2.22	2.14	1.86	1.00	1.39	1.19
e_{dist}	20.1	37.4	28.5	53.1	41.3	29.9	23.8	15.7	24.8	16.2
σ_{dist}	10.7	18.6	14.6	24.0	17.9	22.5	10.5	6.3	8.8	6.6

Table 1: Pose estimation accuracy for all considered activities. Differences to ground truth poses are shown as joint angles (e_{ang} in degrees per joint) and as distances (e_{dist} in millimeters per joint), averaged over all experiments. Standard deviations σ_{ang} and σ_{dist} are provided.

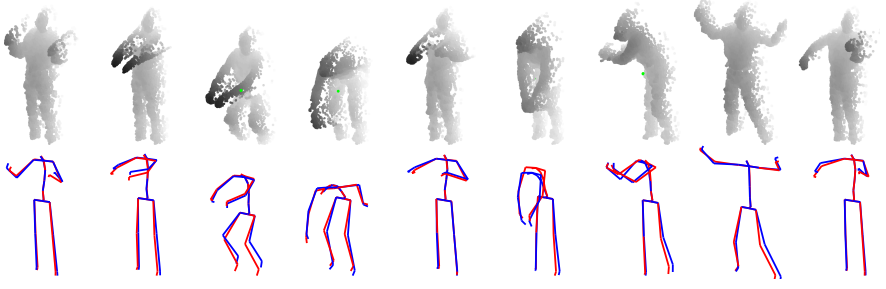

Figure 6: Illustration of pose estimation accuracy for sample frames of different activities. *Top:* Segmented input 3D ToF data of a person. *Bottom:* Corresponding estimated (red) and ground truth (blue) full-body poses.

and the ground truth. Averaged over all frames of the testing sequences, we achieved $\bar{e}_{\text{ang}} = 4.21^\circ$ per joint and $\bar{e}_{\text{dist}} = 29.1$mm. As shown in Table 1, the deviation from the ground truth only increases for fast movements with a large variability, such as *jumping jack*. Figure 6 illustrates the achievable accuracy qualitatively. Our results are comparable to other state-of-the-art methods using visual observations as input [1, 24].

4 Discussion and Conclusion

The proposed motion analysis approach is intended for human-machine interaction applications, where a set of commands is defined in advance, allowing for a training phase. The ability of our method to combine activity recognition and full-body tracking permits not only triggering discrete commands with specific movements, but also performing fine-grained control, e.g based on the exact pose and speed of execution. The introduced learned model integrates prior knowledge on human motion, making it possible to handle the non-trivial multi-activity tracking problem efficiently while relying on simple depth cues.

The depth images provided by ToF cameras are comparable to low-resolution stereo images without the computational complexity; they allow us to easily segment the foreground and to extract features that are invariant to scale and translation. We do not explicitly address rotational invariance. This is sufficient for interactions where the person is facing the camera. Although the sequences in the experiments included small variations, more severe rotations of the person would require to include lateral poses in the training data. Experiments using a 2D version of our depth descriptor confirm the advantage of using the 3D information, as well as the ability of our simple depth descriptor to capture the 3D information. As expected, lower recognition rates are obtained ($\sim 10\%$ less) with confusions occurring especially be-

tween movements that extend to the third dimension (e.g. *punching*, *clapping*).

We have presented a method for human motion analysis using simple depth cues from ToF cameras. Our learned prior motion model enables simultaneous activity recognition and full-body pose tracking. The method is efficient, since we track poses in a low-dimensional space of manifold embeddings and use non-linear regression to relate the embedding space to observations and to full-body poses. Experiments show that the method can reliably recognize movements of multiple activities and estimate full-body pose to a high precision.

Acknowledgements This work was partially supported by the German Federal Ministry of Education and Research (AVILUS project, grant no. 01 IM 08 001 A).

References

[1] A Agarwal, B Triggs, and F Montbonnot. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28 (1):44–58, 2006.

[2] A O Balan, L Sigal, M J Black, J E Davis, and H W Haussecker. Detailed human shape and pose from images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2007.

[3] M Belkin and P Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373 – 1396, Feb 2003.

[4] A Elgammal and C Lee. The role of manifold learning in human motion analysis. *Human Motion Understanding, Modeling, Capture and Animation*, pages 1–29, 2008.

[5] V Ganapathi, C Plagemann, D Koller, and S Thrun. Real time motion capture using a single time-of-flight camera. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[6] M Haker, M Böhme, M Martinetz, and E Barth. Scale-invariant range features for time-of-flight camera applications. *ToF Camera Based Computer Vision*, 2008.

[7] M B Holte, T B Moeslund, and P Fihl. Fusion of range and intensity information for view invariant gesture recognition. *Computer Vision and Pattern Recognition Workshops*, May 2008.

[8] M Isard and A Blake. Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision*, Jan 1998.

[9] Michael Isard and Andrew Blake. A mixed-state condensation tracker with automatic model-switching. *IEEE International Conference on Computer Vision (ICCV)*, pages 107–112, 1998.

[10] T Jaeggli, E Koller-Meier, and L Van Gool. Learning generative models for multi-activity body pose estimation. *International Journal of Computer Vision*, 83(2):121–134, 2009.

[11] R Jensen, R Paulsen, and R Larsen. Analyzing gait using a time-of-flight camera. *Scandinavian Conference on Image Analysis*, pages 21–30, 2009.

[12] A E Johnson and M Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, May 1999.

[13] A Kanaujia, C Sminchisescu, and D Metaxas. Spectral latent variable models for perceptual inference. *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[14] R Koch, I Schiller, B Bartczak, F Kellner, and K Köser. MixIn3D: 3D mixed reality with tof-camera. *Dynamic 3D Imaging*, pages 126–141, 2010.

[15] A Kolb, E Barth, R Koch, and R Larsen. Time-of-flight sensors in computer graphics. *EUROGRAPHICS*, pages 119–134, 2009.

[16] E Kollorz, J Penne, J Hornegger, and A Barke. Gesture recognition with a time-of-flight camera. *International Journal of Intelligent Systems Technologies and Applications*, 5 (3):334–343, 2008.

[17] Z Lu, M Carreira-Perpinan, and C Sminchisescu. People tracking with the laplacian eigenmaps latent variable model. *Neural Information Processing Systems (NIPS)*, 2007.

[18] C Plagemann, V Ganapathi, and D Koller. Real-time identification and localization of body parts from depth images. *IEEE International Comference on Robotics and Automation (ICRA)*, Jan 2010.

[19] J Rodgers, D Anguelov, H.-C Pang, and D Koller. Object pose detection in range scan data. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[20] L Schwarz, D Mateus, and N Navab. Multiple-activity human body tracking in unconstrained environments. *Articulated Motion and Deformable Objects (AMDO)*, 2010.

[21] S Soutschek, J Penne, J Hornegger, and J Kornhuber. 3D gesture-based scene navigation in medical imaging applications using time-of-flight cameras. *Computer Vision and Pattern Recognition Workshops*, Apr 2008.

[22] R Urtasun, DJ Fleet, A Hertzmann, and P Fua. Priors for people tracking from small training sets. *IEEE International Conference on Computer Vision (ICCV)*, 2005.

[23] R Urtasun, DJ Fleet, and P Fua. 3D people tracking with gaussian process dynamical models. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[24] D Vlasic, R Adelsberger, G Vannucci, and J Barnwell. Practical motion capture in everyday surroundings. *ACM Transactions on Graphics (TOG)*, 2007.

[25] D Weinland, R Ronfard, and E Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):249–257, Nov 2006.

[26] D Weinland, E Boyer, and R Ronfard. Action recognition from arbitrary views using 3D exemplars. *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[27] Y Zhu, B Dariush, and K Fujimura. Controlled human pose estimation from depth image streams. *Computer Vision and Pattern Recognition Workshops*, 2008.